National Academy of Sciences Engineering and Medicine Space Weather II Workshop

Data Science and Analytics: Machine Learning and Validation Panel

April 13, 2022

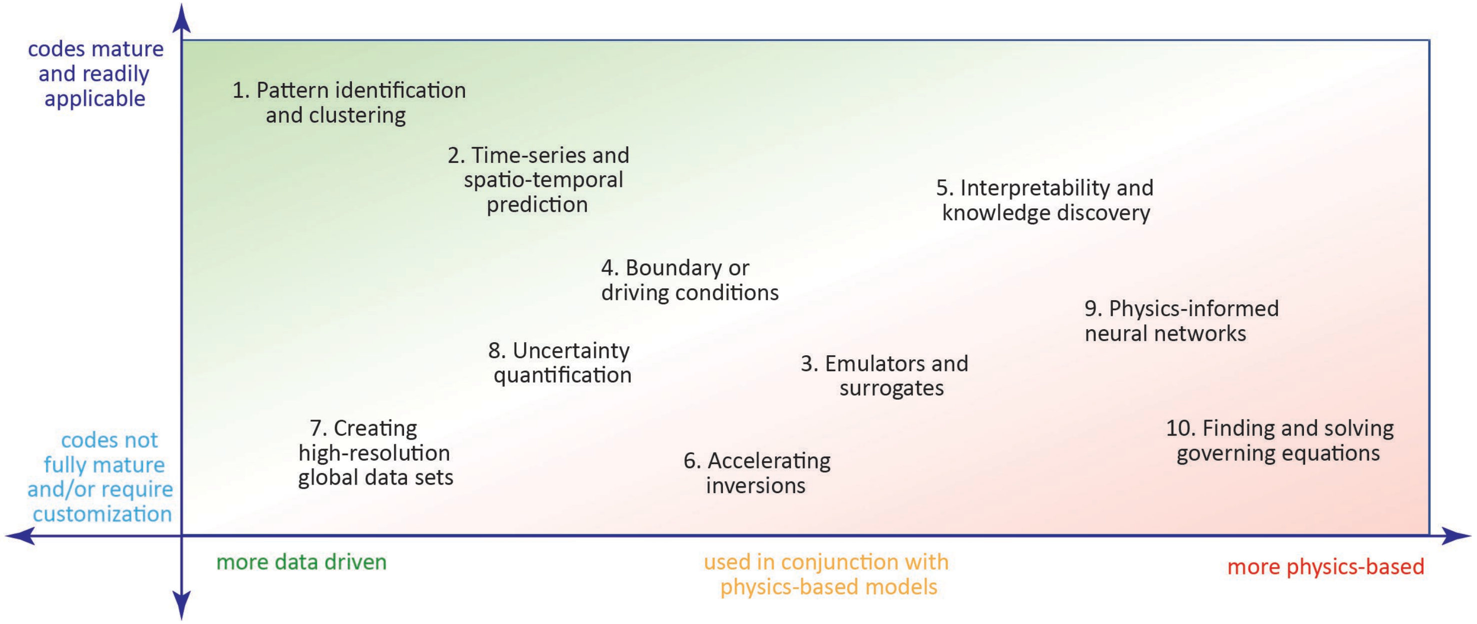
Asti Bhatt, Shasha Zou, Morris Cohen, David Fouhey and Hannah Marlowe, Jacob Bortnik

The National Academies of SCIENCES ENGINEERING MEDICINE

Extracting knowledge from data: machine learning in the physical sciences

- Jacob Bortnik: UCLA professor, space physics + machine learning
- Machine Learning (ML) in general applications:
    1. Obtain data -> build a model
    2. Test on "unseen" data (hold out test & validation set)
    3. Make forecasts/predictions
    4. Adjust model as needed.
    5. Model generally "black box", hard to know what its doing
- Machine Learning for physical sciences (e.g., space weather)
    1. Build a model that predicts well (as above)
    2. Extract physical insight/understanding
        - Physical laws are obeyed: ensure symmetries, invariants, conserved quantities
        - Interpretability: which inputs controlled the output? Globally? Locally?
        - Causality flow: how did information get transmitted in the system? What controls what?
        - Extract governing equations: what are we missing in our understanding.

Extracting knowledge from data: AI/ML for space science, space weather



Bortnik, J., and E. Camporeale (2021), Ten ways to apply machine learning in Earth and space sciences, *Eos, 102*, https://doi.org/10.1029/2021EO160257.

- Space scientist of tomorrow need to have a working knowledge of ML
  - Volumes of data are growing, can't do science the traditional way
  - ML supersedes physics-based models in many cases
  - Students need to build ML models themselves
  - Students need to understand and converse with deep ML specialists
- Needs:
  - Curriculum development, books, classes
  - Workshops/meetings
  - Algorithmic development: "ML for physics" algorithms are only a few years old, need R&D.

SPACE PHYSICS DATA ECOSYSTEM FOR DATA SCIENCE APPLICATIONS

Asti Bhatt

Ionosphere-Thermosphere research scientist

Data provider

Machine-learning dabbler

SRI International

CHALLENGES OF USING BIG DATA METHODS ON SPACE PHYSICS DATA

Machine learning methods have been in existence for longer than the preponderance of Big datasets.

History suggests that creation of **labeled data** (often called **ML-ready**) sparked the adoption and use of the machine-learning techniques for mainstream (e.g., MNIST/ImageNet) and scientific (e.g., Solar flares) use

Do we have the needed data ecosystem to create ML-ready datasets?

Space physics data infrastructure, especially for the ground-based observations is severely lacking in its ability to provide standardized data products.

Findable - Many

Accessible - Some, but process not smooth

Interoperable - With a ton of effort (individual case studies)

Reusable - No

WHAT DO WE NEED TO DO?

Data providers:

- Standardized metadata
- Adoption of FAIR practices

Funding agencies support:

- Support for data providers to do the above
- Support for centralized data infrastructure
- Support for human resources needed for software development
- Help community adapt reproducible computational research practices, which includes -
  - open data
  - open-source software
  - centralized persistent data identifiers
  - use of containers for publishing results

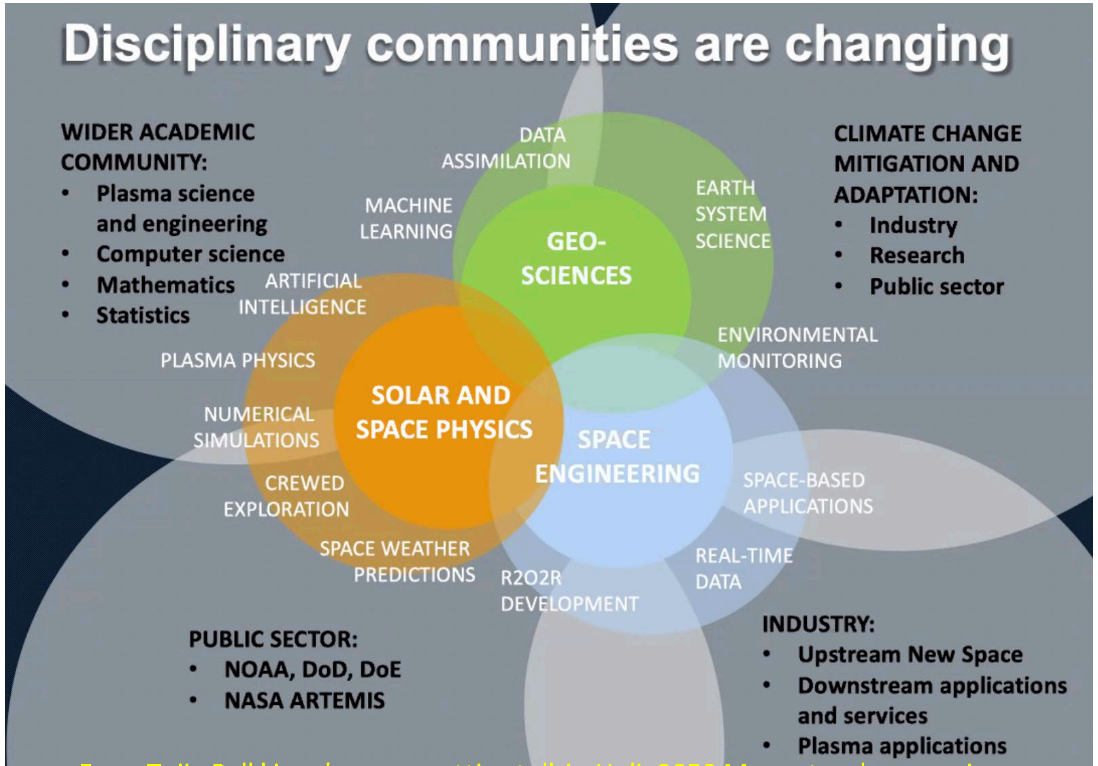
Space Weather & Machine Learning
When Disciplinary Boundaries are Melting

Shasha Zou
University of Michigan

1. What Investments are needed to produce physics-informed machine learning for space weather?

Sustained yet agile research funding programs needed to support interdisciplinary research teams.

Interdisciplinary academic programs/courses needed to train next-generation space physicists equipped with ML knowledge and prepare ML experts with necessary domain knowledge



From Tuija Pulkkinen's scene setting talk in Helio2050 Magnetosphere session

2. Are there adequate data sources and curation for machine learning use/validation/uncertainty determination?

Still data-sparse due to disparate spatial and temporal scales of space weather phenomena

Satellite constellations needed to provide multi-point observations in diverse space plasma regimes

Closer collaboration between government agencies and private sector needed to harness the boom of private space industry and AI technologies

Morris Cohen

Space Weather Operations and Research Infrastructure Workshop

Phase II

13-Apr-2022

Data science and analytics: Machine learning and validation panel

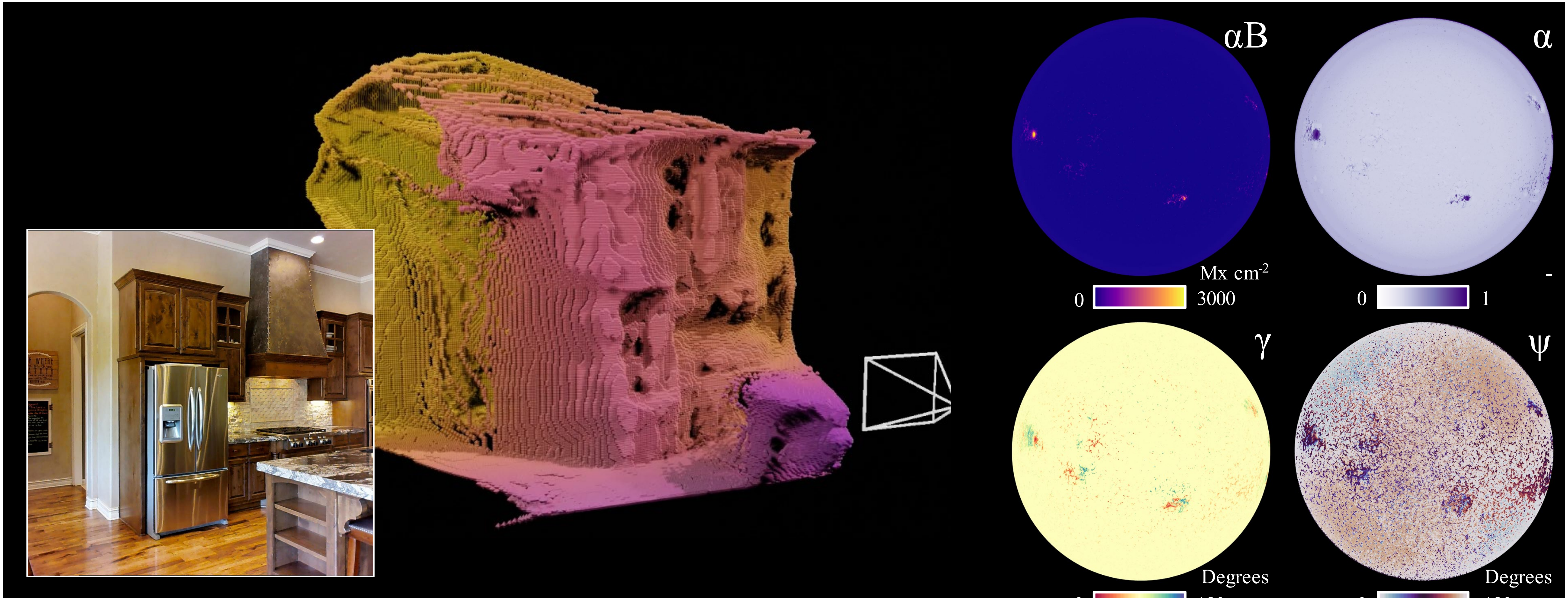
ML and space weather is a two-way street

- We are tackling difficult and meaty data science problems in the field of space sciences.
  - Big data, misbehaved/missing/inconsistent data, need for explainability, fusing theory with models, merging datasets that span decades but do not overlap, etc.
- Let's not just "bring in data science", but market the potential for new advances *to data science*
  - This will in turn impact other research areas
- A few ideas
  - Incentivize space-based data science course development at computing programs and departments
  - Fund new faculty hires of space science PhD in computing departments
  - Targeted funding programs in computing-centric directorates centered on "misbehaved data", using space science as a sandbox

Ask not what ~~your country~~ *machine learning*
can do for ~~you~~ *space weather*,
but what ~~you~~ *space weather*
can do for ~~your country~~ *machine learning*

Maximizing the Impact of Core ML+CV in SW

David Fouhey, Assistant Professor of EECS, University of Michigan EECS



Left: Kulkarni, Johnson, Fouhey. Arxiv. Right: Higgins, Fouhey, et al. *SynthIA*, ApJS.

Maximizing the Impact of Core ML+CV in SW

Getting Core ML People Started	Building Expertise	Steady-State Involvement	Future Possibilities
Space Weather (SW) competes with other areas for ML attention.	Progress is hard for Core MLers due to knowledge gaps, many gotchas.	Core MLers can't find SW problems; areas are often terra incognita.	**Imagine**: (1) data pipelines & instruments that let advances in ML aid SW; (2) work that advances both SW + ML; (3) the co-design of both instruments and ML solutions.
Need: easy entry point. What I found helpful: NASA FDL; too many people to name; pre-picked problems with *known* impact.	**Need**: welcome mat. Continuing funding & support for: data products; software to avoid trivial bugs; people who collab with people like me.	**Need**: quality SW & ML work & validation. SW & ML novelty is bonus; first need baselines. Common Task helps, has perils.	***Let's get to this future!***

HANNAH MARLOWE

END

Panel structure

- KD Leka moderate and introduce panelists

Wednesday

2:00 pm **Data science and analytics: Machine learning and validation panel**

<u>Moderator:</u> *KD Leka*, Committee Member

Jacob Bortnik, UCLA

Asti Bhatt, SRI

Shasha Zou, University of Michigan

Morris Cohen, Georgia Tech

David Fouhey, University of Michigan

Hannah Marlowe, Amazon Web Services

2:45 pm **Break**

Space Weather Operations and Research Infrastructure Workshop: Phase II	
Time (EST)	**Session title**
Monday April 11	
11.00-11.30	Keynote
11.30-12.15	Agency panel: Recent updates
	BREAK
13.00-13.45	Interagency partnerships panel: New ways of working
13.45-14.15	Space weather operations panel
14.15-14.45	People perspectives panel
	BREAK
15.30-16.50	Keynote presentations: New research needs (4x20 min)
16.50-17.20	Keynote presentation: Prediction of ground effects
Tuesday April 12	
11.00-11.40	Research needs: Ionospheric state and irregularities panel
11.40-12.20	Research needs: Cross scale and cross-region coupling panel
	BREAK
13.00-13.45	Observation & Model needs: Sun panel
13.45-14.30	Observation & Model needs: Solar Wind panel
14.30-15.15	Observation & Model needs: Magnetosphere panel
	BREAK
16.00-16.45	Observation & Model needs: Ionosphere and thermosphere panel
16.45-17.30	Observation & Model needs: Ground effects panel
Wednesday April 13	
11.00-11.45	Data Science and Analytics: Keynotes
11.45-12.30	Data Science and Analytics: Data/model resources and curation panel
	BREAK
13.15-14.00	Data Science and Analytics: Data fusion and assimilation panel
14.00-14.45	Data Science and Analytics: Machine learning and validation panel
	BREAK
15.30-17.00	New Architectures: Solar and heliosphere panel
17.00-18.00	Poster Session

Background and Interests

- Associate Professor, Electrical and Computer Engineering, Georgia Tech
- 2012-2013 AAAS Science and Technology Policy Fellow, served at the NSF
- PhD background: Ionospheric remote sensing with VLF (1-50 kHz) radio waves
- More recent research areas: Novel time-varying antennas, power grid cybersecurity, machine learning in space weather
- Hobbies: Building wood-fired pizza ovens, and then cooking in them